Generalizing an Approximate Subgraph Matching-based System to Extract Events in Molecular Biology and Cancer Genetics

Haibin Liu
haibin.liu@nih.gov
NCBI, Bethesda, MD, USA

Karin Verspoor
karin.verspoor@nicta.com.au
NICTA, Melbourne, VIC, Australia

Donald C. Comeau
comeau@ncbi.nlm.nih.gov
NCBI, Bethesda, MD, USA

Andrew MacKinlay
andrew.mackinlay@nicta.com.au
NICTA, Melbourne, VIC, Australia

W. John Wilbur
wilbur@ncbi.nlm.nih.gov
NCBI, Bethesda, MD, USA

Abstract

We participated in the BioNLP 2013 shared tasks, addressing the GENIA (GE) and the Cancer Genetics (CG) event extraction tasks. Our event extraction is based on the system we recently proposed for mining relations and events involving genes or proteins in the biomedical literature using a novel, approximate subgraph matching-based approach. In addition to handling the GE task involving 13 event types uniformly related to molecular biology, we generalized our system to address the CG task targeting a challenging set of 40 event types related to cancer biology with various arguments involving 18 kinds of biological entities. Moreover, we attempted to integrate a distributional similarity model into our system to extend the graph matching scheme for more events. In addition, we evaluated the impact of using paths of all possible lengths among event participants as key contextual dependencies to extract potential events as compared to using only the shortest paths within the framework of our system.

We achieved a 46.38% F-score in the CG task and a 48.93% F-score in the GE task, ranking 3rd and 4th respectively. The consistent performance confirms that our system generalizes well to various event extraction tasks and scales to handle a large number of event and entity types.

1 Introduction

Understanding the sophisticated interactions between various components of biological systems and consequences of these biological processes on the function and behavior of the systems provides profound impacts on translational biomedical research, leading to more rapid development of new therapeutics and vaccines for combating diseases. For the past five years, the BioNLP shared task series has served as an instrumental platform to promote the development of text mining methodologies and resources for the automatic extraction of semantic events involving genes or proteins such as gene expression, binding, or regulatory events from the biomedical literature (Kim et al., 2009; Kim et al., 2011). An event typically captures the association of multiple participants of varying numbers and with diverse semantic roles (Ananiadou et al., 2010). Since events often serve as participants in other events, the extraction of such nested event structures provides an integrated, network view of these biological processes.

Previous shared tasks focused exclusively on events at the molecular and sub-cellular level. However, biological processes at higher levels of organization are equally important, such as cell proliferation, organ growth and blood vessel development. While preserving the classic event extraction tasks such as the GE task, the BioNLP-ST 2013 broadens the scope of application domains by introducing many new issues in biology such as cancer genetics and pathway curation. On behalf of NCBI (National Center for Biotechnology Information), our team participated in the GENIA (GE) task and the Cancer Genetics (CG) task. Compared to the GE task that aims for 13 types of events concerning the protein NF-κB, the CG task targets a challenging set of 40 types of biological processes related to the development and progression of cancer involving 18 entity types. This additionally requires that event extraction systems be able to associate entities and events at the molecular level with anatomy level effects and organism level outcomes of cancer biology.

Our event extraction is based on the system we recently proposed for mining relations and events involving genes or proteins in the biomedical literature using a novel, Approximate Subgraph Matching-based (ASM) approach (Liu et al., 2013a). When evaluated on the GE task of the BioNLP-ST 2011, its performance is comparable to the top systems in extracting 9 types of biological events. In the BioNLP-

Proceedings of the BioNLP Shared Task 2013 Workshop, pages 76–85,
Sofia, Bulgaria, August 9 2013.

ST 2013, we generalized our system to investigate both CG and GE tasks. Moreover, we attempted to integrate a distributional similarity model into the system to extend the graph matching scheme for more events. The graph representation that considers paths of all possible lengths (all-paths) between any two nodes has been encoded in graph kernels used in conjunction with Support Vector Machines (SVM), and led to state-of-the-art performance in extracting protein-protein (Airola et al., 2008) and drug-drug interactions (Zhang et al., 2012). Borrowing from the idea of the all-paths representation, in addition, we evaluated the impact of using all-paths among event participants as key contextual dependencies to extract potential events as compared to using only the shortest paths within the framework of our system.

The rest of the paper is organized as follows: In Section 2, we briefly introduce our ASM-based event extraction system. Section 3 describes our experiments aiming to extend our system. Section 4 elaborates some implementation details and Section 5 presents our results and discussion. Finally, Section 6 summarizes the paper and introduces future work.

2 ASM-based Event Extraction

The underlying assumption of our event extraction approach is that the contextual dependencies of each stated biological event represent a typical context for such events in the biomedical literature. Our approach falls into the machine learning category of instance-based reasoning (Alpaydin, 2004). Specifically, the key contextual structures are learned from each labeled positive instance in a set of training data and maintained as event rules in the form of subgraphs. Extraction of events is performed by searching for an approximate subgraph isomorphism between key dependencies and input sentence graphs using an approximate subgraph matching (ASM) algorithm designed for literature-based relational knowledge extraction (Liu et al., 2013a). By introducing error tolerance into the graph matching process, our approach is capable of retrieving events encoded within complex dependency contexts while maintaining the extraction precision at a high level. The ASM algorithm has been released as open source software[1]. See (Liu et al., 2013a) for more details on the ASM algorithm, its complexity and the comparison with existing graph distance metrics.

Figure 1 illustrates the overall architecture of our ASM-based system with three core components highlighted: rule induction, sentence matching and rule set optimization. Our approach focuses on extracting events expressed within the boundaries of a single sentence. It is also assumed that entities involved in the target event have been annotated. Next, we briefly describe the core components of the system.

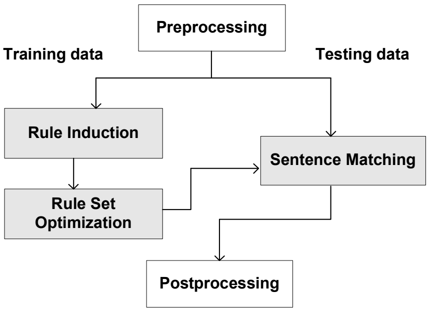

Figure 1: ASM-based Event Extraction Framework

2.1 Rule Induction

Event rules are learned automatically using the following method. Starting with the dependency graph of each training sentence, for each annotated event, the shortest dependency path connecting the event trigger to each event argument in the undirected version of the graph is selected. While additional information such as individual words in each sentence (bag-of-words), sequences of words (n-grams) and semantic concepts is typically used in the state-of-the-art supervised learning-based systems to cover a broader context (Airola et al., 2008; Buyko et al., 2009; Björne et al., 2012), the shortest path between two tokens in the dependency graph is particularly likely to carry the most valuable information about their mutual relationship (Bunescu and Mooney, 2005a; Thomas et al., 2011b; Rinaldi et al., 2010). In case there exists more than one shortest path, all of them are considered. For multi-token event triggers, the shortest path connecting every trigger token to each event argument is extracted, and the union of the paths is then computed for each trigger. For regulatory events that take a sub-event as an argument, the shortest path is extracted so as to connect the trigger of the main event to that of the sub-event.

For complex events that involve multiple arguments, we computed the dependency path union of all shortest paths from trigger to each event argument, resulting in a graph in which all event participants are jointly depicted. Individual dependency paths connecting triggers to each argument are also considered to determine event arguments independently. If the

[1] http://asmalgorithm.sourceforge.net

resulting arguments share the same event trigger, they are grouped together to form a potential event. In our approach, the individual paths aim to retrieve more potential events while the path unions retain the precision advantage of joint inference.

While the dependencies of such paths are used as the graph representation of the event, a detailed description records the participants of the event, their semantic role labels and the associated nodes in the graph. All participating biological entities are replaced with a tag denoting their entity type, e.g. "Protein" or "Organism", to ensure generalization of the learned rules. As a result, each annotated event is generalized and transformed into a generic graph-based rule. The resulting event rules are categorized into different target event types.

2.2 Sentence Matching

Event extraction is achieved by matching the induced rules to each testing sentence and applying the descriptions of rule tokens (e.g. role labels) to the corresponding sentence tokens. Since rules and sentence parses all possess a graph representation, event recognition becomes a subgraph matching problem. We introduced a novel *approximate subgraph matching* (ASM) algorithm (Liu et al., 2013a) to identify a subgraph isomorphic to a rule graph within the graph of a testing sentence. The ASM problem is defined as follows.

Definition 1. An event rule graph $G_r = (V_r, E_r)$ is *approximately isomorphic* to a subgraph S_s of a sentence graph $G_s = (V_s, E_s)$, denoted by $G_r \cong_t S_s \subseteq G_s$, if there is an injective mapping $f : V_r \to V_s$ such that, for a given threshold t, $t \geq 0$, the subgraph distance between G_r and G_s satisfies $0 \leq \text{subgraphDist}_f(G_r, G_s) \leq t$, where $\text{subgraphDist}_f(G_r, G_s) = w_s \times \text{structDist}_f(G_r, G_s) + w_l \times \text{labelDist}_f(G_r, G_s) + w_d \times \text{directionalityDist}_f(G_r, G_s)$.

The subgraph distance is proposed to be the weighted summation of three penalty-based measures for a candidate match between the two graphs. The measure **structDist** compares the distance between each pair of matched nodes in one graph to the distance between corresponding nodes in the other graph, and accumulates the structural differences. The distance in rule graphs is defined as the length of the shortest path between two nodes. The distance in sentence graphs is defined as the length of the path between corresponding nodes that leads to minimum structural difference with the distance in rule graphs. Because dependency graphs are edge-labeled, oriented graphs, the measures **labelDist** and **directionalityDist** evaluate respectively the overall differences in edge labels and directionalities on the compared path between each pair of matched nodes in the two graphs. The real numbers w_s, w_l and w_d are non-negative weights associated with the measures.

The weights w_s, w_l and w_d are defaulted to be equal but can be tuned to change the emphasis of the overall distance function. The distance threshold t controls the isomorphism quality of the retrieved subgraphs from sentences. A smaller t allows only limited variations and always looks for a sentence subgraph as closely isomorphic to the rule graph as possible. A larger t enables the extraction of events described in complicated dependency contexts, thus increasing the chance of retrieving more events. However, it can incur a bigger search cost due to the evaluation of more potential solutions.

An iterative, bottom-up matching process is used to ensure the extraction of complex and nested events. Starting with the extraction of simple events, simple event rules are first matched with a testing sentence. Next, as potential arguments of higher level events, obtained simple events continue to participate in the subsequent matching process between complex event rules and the sentence to initiate the iterative process for detecting complex events with nested structures. The process terminates when no new candidate event is generated for the testing sentence.

During the matching phase we relax the event rules that contain sub-event arguments such that any matched event can substitute for the sub-event. We believe that the contextual structures linking annotated sub-events of a certain type are generalizable to other event types. This relaxation increases the chance of extracting complex events with nested structures but still takes advantage of the contextual constraints encoded in the rule graphs.

2.3 Rule Set Optimization

Typical of instance-based reasoners, the accuracy of rules with which to compare an unseen sentence is crucial to the success of our approach. For instance, a *Transcription* rule encoding a noun compound modification dependency between "TNF" and "mRNA" derived from an event context "expression of TNF mRNA" should not produce a *Transcription* event for the general phrase "level of TNF mRNA" even though they share a matchable dependency. Such matches result in false positive events.

Therefore, we measured the accuracy of each rule r_i in terms of its prediction result via Eq.(1). For rules that produce at least one prediction, we ranked them by $Acc(r_i)$ and excluded the ones with a $Acc(r_i)$ ratio lower than an empirical threshold, e.g. 1:4.

$$Acc(r_i) = \frac{\#correct\_predictions\_by\_r_i}{\#total\_predictions\_by\_r_i} \quad (1)$$

Because of nested event structures, the removal of some rules might incur a propagating effect on rules relying on them to produce arguments for the extraction of higher order events. Therefore, an iterative rule set optimization process, in which each iteration performs sentence matching, rule ranking and rule removal sequentially, is conducted, leading to a converged, optimized rule set. While the ASM algorithm aims to extract more potential events, this performance-based evaluation component ensures the precision of our event extraction framework.

3 Extensions to Event Extraction System

In the BioNLP-ST 2013, we attempted two different ways to extend the current event extraction system: (1) integrate a distributional similarity model into the system to extend the graph matching scheme for more events; (2) use paths of all possible lengths (all-paths) among event participants as key contextual dependencies to extract events. We next elaborate these system extensions in detail.

3.1 Integrating Distributional Similarity Model

The proposed subgraph distance measure of the ASM algorithm focuses on capturing differences in the overall graph structure, edge labels and directionalities. However, when determining the injective node mapping between graphs, the matching remains at the surface word level.

In the current setting, various node features can be considered when comparing two graph nodes, resulting in different matching criteria. The features include POS tags (P), event trigger (T), token lemmas (L) and tokens themselves (A). For instance, a matching criterion, “P*+L”, requires that the relaxed POS tags (P*) and the lemmatized form (L) of tokens be identical for each rule node to match with a sentence node. The relaxed POS allows the plural form of nouns to match with the singular form, and the conjugations of verbs to match with each other. However, the inability to go beyond surface level matching prevents node tokens that share similar meaning but possess distinct orthography from matching with each other. For instance, a mismatch between rule token “crucial” and a sentence token “critical’ could lead to an undiscovered *Positive_regulation* event.

We attempted to use only POS information in the node matching scheme and observed a nearly 14% increase in recall (Liu et al., 2013b). However, the precision drops sharply, resulting in an undesirable F-score. This indicates that the lexical information is a critical supplement to the contextual dependency constraints in accurately capturing events within the framework of our system. Moreover, we attempted to extend the node matching using the synsets of WordNet (Fellbaum, 1998) to allow tokens to match with their synonyms (Liu et al., 2011). However, since WordNet is developed for the general English language, it relates biomedical terms e.g., “expression” with general words such as “aspect” and “face”, thus leading to incorrect events.

In this work, we integrated a distributional similarity model (DSM) into our node matching scheme to further improve the generalization of event rules. A distributional similarity model is constructed based on the distributional hypothesis (Harris, 1954): words that occur in the same contexts tend to share similar meanings. We expect that the incorporation of DSM will enable our system to capture matching tokens in testing sentences that do not appear in the training data while maintaining the extraction precision at a high level. There have been many approaches to compute the similarity between words based on their distribution in a corpus (Landauer and Dumais, 1997; Pantel and Lin, 2002). The output is a ranked list of similar words to each word. We reimplemented the model proposed by (Pantel and Lin, 2002) in which each word is represented by a feature vector and each feature corresponds to a context where the word appears. The value of the feature is the pointwise mutual information (Manning and Schütze, 1999) between the feature and the word. Let c be a context and $F_c(w)$ be the frequency count of a word w occurring in context c. The pointwise mutual information, $mi_{w,c}$ between c and w is defined as:

$$mi_{w,c} = \frac{\frac{F_c(w)}{N}}{\frac{\sum\limits_i F_i(w)}{N} \times \frac{\sum\limits_j F_c(j)}{N}} \quad (2)$$

where $N = \sum\limits_i \sum\limits_j F_i(j)$ is the total frequency count of all words and their contexts.

Since mutual information is known to be biased towards infrequent words/features, the above mutual information value is multiplied by a discounting factor as described in (Pantel and Lin, 2002). The similarity between two words is then computed using the cosine coefficient (Salton and McGill, 1986) of their mutual information vectors.

We experimented with two different approaches to integrate the DSM into our event extraction system. First, the model is directly embedded into the node matching scheme. Once a match cannot be determined by surface tokens, the DSM is invoked to allow a match if the sentence token appears in the list of the top M most similar words to the rule token. Second, additional event rules are generated by replacing corresponding rule tokens with their top M most similar words, rather than allow DSM to participate in the node matching. While the first method measures the consolidated extraction ability of an event rule by combining its DSM-generalized performance, the second approach provides a chance to evaluate the impact of each DSM-introduced similar word individually on event extraction.

3.2 Adopting All-paths for Event Rules

Airola *et al.* proposed an all-paths graph (APG) kernel for extracting protein-protein interactions (PPI), in which the kernel function counts weighted shared dependency paths of all possible lengths (Airola et al., 2008). Thomas *et al.* adopted this kernel as one of the three models used in the ensemble learning for extracting drug-drug interactions (Thomas et al., 2011a) and won the recent DDIExtraction 2011 challenge (Segura-Bedmar et al., 2011). The JULIE lab adapted the APG kernel to event extraction using syntactically pruned and semantically enriched dependency graphs (Buyko et al., 2009).

The graph representation of the kernel consists of two sub-representations: the full dependency parse and the surface word sequence of the sentence where a pair of interacting entities occurs. At the expense of computational complexity, this representation enables the kernel to explore broader contexts of an interaction, thus taking advantage of the entire dependency graph of the sentence. When comparing two interaction instances, instead of using only the shortest path that might not always provide sufficient syntactic information about relations, the kernel considers paths of all possible lengths between any two nodes. More recently, a hash subgraph pairwise (HSP) kernel-based approach was also proposed for drug-drug interactions and adopts the same graph representation as the APG kernel (Zhang et al., 2012).

In contrast, the graph representation that our ASM algorithm searches in a sentence is inherently restricted to the shortest path among target entities in event rules, as described in Section 2.2. Borrowing from the idea of the all-path graph representation, in this work we attempted to explore contexts beyond the shortest paths to enrich our rule set. We evaluated within the framework of our system the impact of using acyclic paths of all possible lengths among event participants as key contextual dependencies to populate the event rule set as compared to using only the shortest paths in the current system setting.

4 Implementation

4.1 Preprocessing

We employed the preprocessed data in the BioC (Comeau et al., 2013) compliant XML format provided by the shared task organizers as supporting resources. The BioC project attempts to address the interoperability among existing natural language processing tools by providing a unified BioC XML format. The supporting analyses include tokenization, sentence segmentation, POS tagging and lemmatization. Different syntactic parsers analyze text based on different underlying methodologies, for instances, the Stanford parser (Klein and Manning, 2003) performs joint inference over the product of an unlexicalized Probabilistic Context-Free Grammar (PCFG) parser and a lexicalized dependency parser while the McClosky-Charniak-Johnson (Charniak) parser (McClosky and Charniak, 2008) is based on N-best parse reranking over a lexicalized PCFG model. In order to take advantage of multiple aspects of structural analysis of sentences, both Stanford parser and Charniak parser, which are among the best performing parsers trained on the GENIA Treebank corpus, are used to parse the training sentences and produce dependency graphs for learning event rules. Only the Charniak parser is used on the testing sentences in the event extraction phase.

4.2 ASM Parameter Setting

The GE task includes 13 different event types. Since each type possesses its own event contexts, an individual threshold t_e is assigned to each type. Together with the 3 distance function weights w_s, w_l and w_d, the ASM requires 16 parameters for the GE event extraction task. Similarly, the ASM requires 43 parameters to cater to the 40 diverse event types of the CG task. As reported in (Liu et al., 2013a), we used a genetic algorithm (GA) (Cormen et al., 2001) to automatically determine values of the 12 ASM parameters for the 2011 GE task using the training data. We inherited these previously determined parameters and adapted them into the 2013 tasks according to the event type and its argument configuration. For instance, "Pathway" events in the CG task is assigned the same t_e as the "Binding" events in the GE task as they possess similar argument configurations.

Table 1 shows the parameter setting for the 2013 GE task with the equal weights $w_s = w_l = w_d$ constraint. The graph node matching criterion "P*+L" that requires the relaxed POS tags and the token lemmas to be identical is used in the ASM.

Parameter	Value	Parameter	Value
$t_{Gene\_expression}$	8	$t_{Ubiquitination}$	3
$t_{Transcription}$	7	$t_{Binding}$	7
$t_{Protein\_catabolism}$	10	$t_{Regulation}$	3
$t_{Phosphorylation}$	8	$t_{Positive\_regulation}$	3
$t_{Localization}$	8	$t_{Negative\_regulation}$	3
$t_{Acetylation}$	3	w_s	10
$t_{Deacetylation}$	3	w_l	10
$t_{Protein_{m}odification}$	3	w_d	10

Table 1: ASM parameter setting for the 2013 GE task

4.3 Distributional Similarity Model

In our implementation, we made following improvements to the original Pantel model (Pantel and Lin, 2002): (1) lemmas of words generated by the BioLemmatizer (Liu et al., 2012) are used to achieve generalization. The POS information is combined with each lemmatized word to disambiguate its category. (2) instead of the linear context where a word occurs, we take advantage of dependency contexts inferred from dependency graphs. For instance, "toxicity→amod" is extracted as a feature of the token "nonhematopoietic JJ". It captures the dependent token, the type and the directionality of the dependency. (3) the resulting $mi_{w,c}$ is scaled into the [0, 1] range by $\frac{\lambda \cdot mi_{w,c}}{1 + \lambda \cdot mi_{w,c}}$ to avoid greater $mi_{w,c}$ values dominating the similarity calculation between words. An empirical $\lambda = 0.01$ is used. (4) while only the immediate dependency contexts of a word are used in our model, our implementation is flexible so that contexts of various dependency depths could be taken into consideration.

In order to cover a wide range of words and capture the diverse usages of them in biomedical texts, instead of resorting to an existing corpus, our distributional similarity model is built based on a random selection of 5 million abstracts from the entire PubMed. When computing $mi_{w,c}$, we filtered out contexts of each word where the word occurs less than 5 times. Eventually, the model contains 2.8 million distinct tokens and 0.4 million features. When it is queried with an amino acid, e.g, "lysine", the top 15 tokens in the resulting ranked list are all correct amino acid names.

5 Results and Discussion

This section reports our results on the GE and the CG tasks respectively, including the attempted extensions to our ASM-based event extraction system.

5.1 GE task

5.1.1 Datasets

The 2013 GE task dataset is composed of full-text articles from PubMed Central, which are divided into smaller segments by the task organizers according to various sections of the articles. Table 2 presents some statistics of the GE dataset.

Attributes Counted	Training	Development	Testing
Full article segments	222	249	305
Proteins	3,571	4,138	4,359
Annotated events	2,817	3,199	3,301

Table 2: Statistics of BioNLP-ST 2013 GE dataset

As distributed, the development set is bigger than the training set. For better system generalization, we randomly reshuffled the data and created a 353/118 training/development division, a roughly 3:1 ratio consistent with the settings in previous GE tasks. The results reported on the training/development data thereafter are based on our new data partition.

5.1.2 GE Results on Development Set

Table 3 shows the event extraction results on the 118 development documents based on event rules derived from different parsers. Only the numbers of unique, optimized rules are reported and those that possess isomorphic graph representations determined by an Exact Subgraph Matching (ESM) algorithm (Liu et al., 2013b) are removed. The ensemble rule set combines rules derived from both parsers and achieves a better performance than that of using individual parsers. It makes sense that the Charniak parser is favored and leads to a performance close to the ensemble performance because sentences from which events are extracted are parsed by the Charniak parser as well. However, we retained the additional rules from the Stanford parser in the hope that they may contribute to the testing data.

When embedding the distributional similarity model (DSM) directly into the graph node matching

Parser Type	Event Rule	Recall	Precision	F-score
Charniak	2,923	47.01%	66.01%	54.91%
Stanford	3,305	43.66%	67.67%	53.08%
Ensemble	4,617	47.45%	65.65%	55.09%

Table 3: Performance of using different parsers

scheme, we performed the DSM on all rule tokens except biological entities, meaning that for each rule token, if a match will be granted if a rule token appears in the top M most similar word list of a sentence token, e.g., "DSM_3" denotes the top 3 similar words determined by the DSM. We further performed DSM only on trigger tokens for comparison, as presented in Table 4.

All Tokens	Recall	Precision	F-score
DSM_1	47.98%	52.56%	50.17%
DSM_3	48.68%	35.07%	40.77%
DSM_10	53.43%	19.38%	28.44%
Trigger Tokens	Recall	Precision	F-score
DSM_1	48.06%	54.22%	50.95%
DSM_3	48.59%	37.00%	42.01%
DSM_10	53.35%	24.65%	33.72%

Table 4: Performance of integrated DSM

Even though the DSM helps to substantially increase the recall to 53.43%, we observed a significant precision drop which leads to an inferior F-score to the ensemble baseline in Table 3. A close evaluation of the generated graph matches reveals that antonyms produced by the DSM contributes to most of the false positive events. For instance, the most similar words for the verb "increase" and the adjective "high" returned by the model are "decrease" and "low" because they tend to occur in the same contexts. Further investigation is needed to automatically filter out the antonyms. When generating additional rules using the top M most similar words from the DSM, since all the rules undergo the optimization process, the event extraction precision is ensured. However, the recall increase from simple events is diluted by the counter effect of the introduced false positives in detecting regulation-related complex events, resulting in a comparable performance to the baseline.

Table 5 gives the performance comparison of using all-paths and the shortest paths in our event extraction system. Using all-paths does not bring in a significant improvement in F-score but takes 27 iterations to optimize as compared to the 5-iteration optimization on shortest paths. Most of the rules induced from all-paths are eventually discarded by the optimization process. The all-paths graph representation was motivated by the observation that shortest paths between candidate entities often exclude relation-signaling words when detecting binary relationships (Airola et al., 2008). Exploring broader contexts ensures such words to be considered. In the event extraction task, however, since triggers have been annotated, they are naturally incorporated into the shortest paths connecting trigger to each event argument. This in part explains why contexts beyond shortest paths did not bring in an appreciable benefit.

All Tokens	Recall	Precision	F-score
All-paths	48.77%	64.64%	55.59%
Shortest paths	47.45%	65.65%	55.09%

Table 5: Performance of using all-paths

5.1.3 GE Results on Testing Set

Since integrating the DSM and all-paths do not provide significant performance improvements to our system, we decided to retain the original settings in the ASM when extracting events from the testing data. While most of the 2011 shared task datasets are composed of PubMed abstracts compared to full-text articles in the 2013 GE task, our system focuses on extracting events expressed within the boundaries of a single sentence. Therefore, in order to take advantage of existing annotated resources, we incorporated the annotated data of 2011 GE task and EPI (Epigenetics and Post-translational Modifications) task to enrich the training instances of corresponding event types of the 2013 GE task. Eventually, we obtained a total of 14,448 rules of different event types from our training data. In practice, it takes the ASM less than a second to match the entire rule set with one document and return results.

Our submitted system achieves a 48.93% F-score on the 305 testing documents of the GE task, ranking 4th among 12 participating teams. Table 6 presents the performance of the top eight systems.

System	Recall	Precision	F-score
EVEX	45.44%	58.03%	50.97%
TEES 2.1	46.17%	56.32%	50.74%
BioSEM	42.47%	62.83%	50.68%
NCBI	40.53%	61.72%	48.93%
DlutNLP	40.81%	57.00%	47.56%
HDS4NLP	37.11%	51.19%	43.03%
NICTANLM	36.99%	50.68%	42.77%
USheff	31.69%	63.28%	42.23%

Table 6: Performance of top 8 systems in GE task

Our performance is within a reasonable margin from the best-performing system "EVEX", and shows an overall superior precision over most participating teams; only two of the top 5 systems obtained a precision in the 60% range. Particularly for the regulation-related complex events, we are the only team that achieved a precision over 55% among all 12 participating systems. This indicates that event rules automatically learned and optimized over training data generalize well to the unseen text, and have the ability to identify precisely corresponding events.

We further evaluated the impact of the additonal training instances from 2011 tasks and the ensemble rule set derived from different parsers as presented in Table 7. With the help from the 2011 data, our F-score is increased by 3% and we became the only team that detected "Ubiquitination" events from testing data. In addition, rules derived from the Stanford parser do not provide additional benefits on the testing data compared to using the Charniak parser alone.

System Attribute	Recall	Precision	F-score
Ensemble 2013 + 2011 data	40.53%	61.72%	48.93%
Ensemble 2013 data	35.63%	63.91%	45.75%
Charniak 2013 data	35.29%	65.71%	45.92%

Table 7: Impact of 2011 data and ensemble rule set

5.2 CG task

5.2.1 Datasets

The CG task dataset is prepared based on a previously released corpus of angiogenesis domain abstracts (Wang et al., 2011). It targets a challenging set of 40 types of biological processes related to the development and progression of cancer involving 18 entity types (Pyysalo et al., 2012). Table 8 presents some statistics of the CG dataset.

Attributes Counted	Training	Development	Testing
Abstracts	300	100	200
Entities	10,935	3,634	6,955
Annotated events	8,803	2,915	5,972

Table 8: Statistics of BioNLP-ST 2013 CG dataset

5.2.2 CG Results on Testing Set

We generalized our event extraction system to the CG task and the corresponding annotated data of the 2011 tasks is also incorporated in the training phase to obtain the optimized event rule set. Due to time constraints, the impact of integrating the DSM and all-paths is not evaluated on the CG task. We achieved a 46.38% F-score on the 200 testing documents of the CG task, ranking 3rd among the 6 participating teams. Table 9 gives the primary evaluation results of the 6 participating teams; only "TEES-2.1" and we participated in both GE and CG tasks. The detailed results of each of the targeted 40 event types is available from the official CG task website.

Team	Recall	Precision	F-score
TEES-2.1	48.76%	64.17%	55.41%
NaCTeM	48.83%	55.82%	52.09%
NCBI	38.28%	58.84%	46.38%
RelAgent	41.73%	49.58%	45.32%
UET-NII	19.66%	62.73%	29.94%
ISI	16.44%	47.83%	24.47%

Table 9: Performance of all systems in 2013 CG task

Inconsistent with other biological entities, the entity annotation for the optional "Site" argument involved in events such as "Binding", "Mutation" and "Phosphorylation" are not provided by the task organizers. We consider that detecting "Site" entities is related to entity detection and we would like to focus our system on the event extraction itself. Thus, we decided to ignore the "Site" argument in our system. However, a problem will arise that even though the other arguments are correctly identified for an event, it might still be evaluated as false positive if a "Site" argument is not detected. This results in both false positive and false negative events. In addition, since we did not perform the secondary task which requires us to detect modifications of the predicted events, including negation and speculation, about 7.5% annotated instances in the testing data are thus missed, causing damage to our recall in the overall evaluation. The organizers have agreed to issue an additonal evaluation that will focus on core event extraction targets excluding optional arguments such as "Site" and the secondary task. We will conduct more detailed analysis on the results once they are made available.

6 Conclusion and Future Work

In the BioNLP-ST 2013, we generalized our ASM-based system to address both GE and CG tasks. We attempted to integrate a distributional similarity model into our system to extend the graph matching scheme. We also evaluated the impact of using paths of all possible lengths among event participants as key contextual dependencies to extract potential events as compared to using only the shortest paths within the framework of our system.

We achieved a 46.38% F-score in the CG task and a 48.93% F-score in the GE task, ranking 3rd and 4th respectively. While the distributional similarity model did not improve the overall performance of our system in the tasks, we would like to further investigate the antonym problem introduced by the model in our future work.

Acknowledgments

This research was supported by the Intramural Research Program of the NIH, NLM.

References

Antti Airola, Sampo Pyysalo, Jari Björne, Tapio Pahikkala, Filip Ginter, and Tapio Salakoski1. 2008. All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. *BMC Bioinformatics*, 9 Suppl 11:s2.

Ethem Alpaydin. 2004. *Introduction to Machine Learning*. MIT Press.

Sophia Ananiadou, Sampo Pyysalo, Jun'ichi Tsujii, and Douglas B. Kell. 2010. Event extraction for systems biology by text mining the literature. *Trends in Biotechnology*, 28(7):381–390.

Jari Björne, Filip Ginter, and Tapio Salakoski. 2012. University of turku in the BioNLP'11 shared task. *BMC Bioinformatics*, 13 Suppl 11:S4.

Razvan C. Bunescu and Raymond J. Mooney. 2005a. A shortest path dependency kernel for relation extraction. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 724–731.

Razvan C. Bunescu and Raymond J. Mooney. 2005b. Subsequence kernels for relation extraction. In *Proceedings of the 19th Conference on Neural Information Processing Systems (NIPS)*. Vancouver, BC, December.

Ekaterina Buyko, Erik Faessler, Joachim Wermter, and Udo Hahn. 2009. Event extraction from trimmed dependency graphs. In *BioNLP '09: Proceedings of the Workshop on BioNLP*, pages 19–27, Morristown, NJ, USA. Association for Computational Linguistics.

Donald C. Comeau, Rezarta Islamaj Doğan, Paolo Ciccarese, Kevin Bretonnel Cohen, Martin Krallinger, Florian Leitner, Zhiyong Lu, Yifan Peng, Fabio Rinaldi, Manabu Torii, Alfonso Valencia, Karin Verspoor, Thomas C. Wiegers, Cathy H. Wu, and W. John Wilbur. 2013. BioC: A minimalist approach to interoperability for biomedical text processing. submitted.

Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. 2001. *Introduction to Algorithms*. The MIT Press.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Zellig Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of BioNLP'09 shared task on event extraction. In *Proceedings of BioNLP Shared Task 2009 Workshop*, pages 1–9. Association for Computational Linguistics.

Jin-Dong Kim, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, Ngan Nguyen, and Jun'ichi Tsujii. 2011. Overview of BioNLP shared task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 1–6. Association for Computational Linguistics, June.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 423–430. Association for Computational Linguistics.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211–240.

Haibin Liu, Ravikumar Komandur, and Karin Verspoor. 2011. From graphs to events: A subgraph matching approach for information extraction from biomedical text. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 164–172. Association for Computational Linguistics, June.

Haibin Liu, Tom Christiansen, William A Baumgartner, and Karin Verspoor. 2012. Biolemmatizer: a lemmatization tool for morphological processing of biomedical text. *Journal of Biomedical Semantics*, 3:3.

Haibin Liu, Lawrence Hunter, Vlado Keselj, and Karin Verspoor. 2013a. Approximate subgraph matching-based literature mining for biomedical events and relations. *PLOS ONE*, 8:4 e60954.

Haibin Liu, Vlado Keselj, and Christian Blouin. 2013b. Exploring a subgraph matching approach for extracting biological events from literature. *Computational Intelligence*.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, USA.

David McClosky and Eugene Charniak. 2008. Self-training for biomedical parsing. In *Proceedings of the Association for Computational Linguistics*, pages 101–104, Columbus, Ohio. The Association for Computer Linguistics.

Patrick Pantel and Dekang Lin. 2002. Discovering word senses from text. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pages 613–619, New York, NY, USA. ACM.

Sampo Pyysalo, Tomoko Ohta, Makoto Miwa, Han-Cheol Cho, Jun'ichi Tsujii, and Sophia Ananiadou. 2012. Event extraction across multiple levels of biological organization. *Bioinformatics*, 28:i575–i581.

Fabio Rinaldi, Gerold Schneider, Kaarel Kaljurand, Simon Clematide, Thrse Vachon, and Martin Romacker. 2010. Ontogene in BioCreative II.5. *IEEE/ACM Trans. Comput. Biology Bioinform.*, 7(3):472–480.

Gerard Salton and Michael J. McGill. 1986. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA.

Isabel Segura-Bedmar, Paloma Martinez, and Daniel Sanchez-Cisneros. 2011. The 1st DDIExtraction-2011 Challenge Task: Extraction of Drug-Drug Interactions from Biomedical Texts. In *Proceedings of the 1st Challenge Task on Drug-Drug Interaction Extraction 2011*, pages 1–9.

Philippe Thomas, Mariana Neves, Illes Solt, Domonkos Tikk, and Ulf Leser. 2011a. Relation extraction for drug-drug interactions using ensemble learning. In *Proceedings of DDIExtraction-2011 challenge task*, pages 11–18.

Philippe Thomas, Stefan Pietschmann, Illés Solt, Domonkos Tikk, and Ulf Leser. 2011b. Not all links are equal: Exploiting dependency types for the extraction of protein-protein interactions from text. In *Proceedings of BioNLP 2011 Workshop*, pages 1–9. Association for Computational Linguistics, June.

Domonkos Tikk, Philippe Thomas, Peter Palaga, Jörg Hakenberg, and Ulf Leser. 2010. A comprehensive benchmark of kernel methods to extract protein–protein interactions from literature. *PLoS Computational Biology*, 6:e1000837, July.

Xinglong Wang, Iain McKendrick, Ian Barrett, Ian Dix, Tim French, Jun'ichi Tsujii, and Sophia Ananiadou. 2011. Automatic extraction of angiogenesis bioprocess from text. *Bioinformatics*, 27(19):2730–2737.

Yijia Zhang, Hongfei Lin, Zhihao Yang, Jian Wang, and Yanpeng Li. 2012. A single kernel-based approach to extract drug-drug interactions from biomedical literature. *PLOS ONE*, 7(11): e48901.